Approaches in Fairness-aware Machine Learning on Various Applications and Evaluations

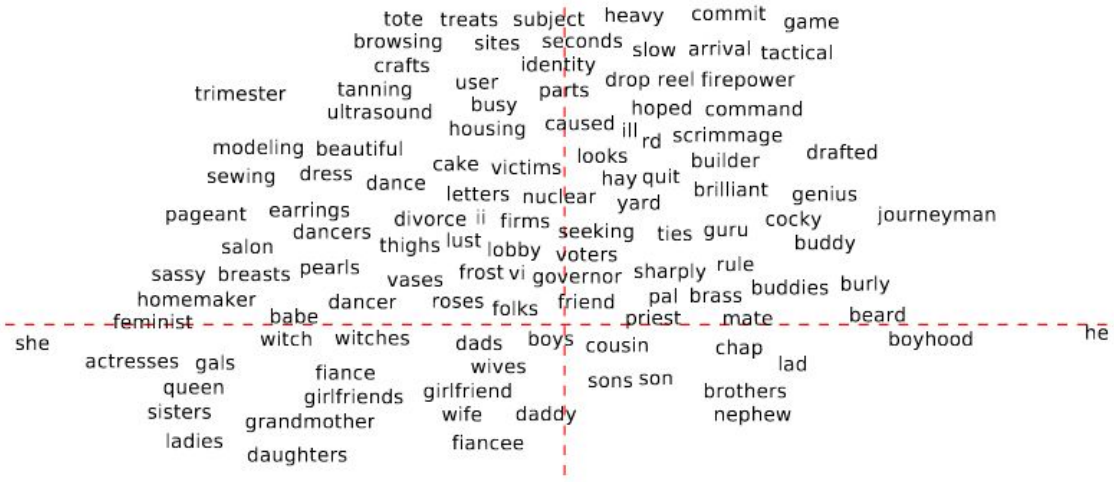
Yoonjin Kim

January 2020

1. Introduction

Machine learning and artificial intelligence are becoming increasingly influential tools for decision making procedures including loan application, job application, college admission, policy making, and decisions that are closely related to our lives. It may seem machine learning is free from discrimination or potentially even a solution for discrimination since there would be no personal bias on the part of a machine, as opposed to a person. However, input data put in machine learning algorithms and automated data analysis to train algorithms are based on the previous dataset which may consist of historical decisions based on discrimination and unfairness. Discrimination is a biased treatment based on their affiliation to different groups(e.g. race) rather than on individual merits[15]. Discrimination is unethical and prohibited by laws including Article 14 of the European Convention on Human Rights (Rome, 4.XI.1950), Title VIII of the Civil Rights Act of 1968 (U.S., Fair Housing Act), and Title IX(U.S. Supreme Court 522-167, 2005)[5]. The European Union established the principle of equal treatment of individuals without discrimination on any ground such as sex, race, colour, language, religion, political or other opinions, national or social origin associated with a national minority, property, birth or other status. In the U.S., the Civil Rights Act of 1968 and 1964 prohibit discrimination in the financing decisions and employment based on race, colour, religion, sex or national origin.

Although the legal regulations against the discrimination are clearly stated, there are cases which show that there are still direct and indirect discriminations such that the EU's Court of Justice ruled that different insurance premiums for women and men constitute sex discrimination and that they are against EU's Charter of Fundamental Rights. Not limited to personal jurisdictions, direct and indirect discriminatory factors are embedded inside input datasets that are used to train learning algorithms. The German credit case study[1] extracted two confidence rules: conf({city=NYC⇒ class=bad})=25% and conf({city=NYC, race=black =>class=bad})=75%. These rules can be translated as "25% of people who live in NYC are assigned the bad credit class" and "75% of black people who live in NYC are assigned bad credit". In this case, it can be inferred that the additional discriminatory item(e.g. race) increased the confidence of the rule of bad credit class up to 75%. Additionally, gender direction on w2vNEWS[7] is one of the salient examples that shows gender discrimination and stereotypes on occupations. The automatically generated analogies for the pair *she-he* using the procedure using *cos* distance between word pair analogies includes nurse-surgeon, interior design-architect, cosmetic-pharmaceuticals, and housewife-shopkeeper.



<Figure 1> Word projected along two axes where the x-axis is a projection onto the difference between word embeddings of the words *he* and *she*, and the y-axis is the direction learned in embedding that captures gender neutral words on I and II Quadrant and gender specific words(e.g. Sister or brother) on III and IV Quadrants. The words above the horizontal line is randomly assigned in this figure[7].

<Figure 1>[7] is a visualization of the subset of words from dataset result of w2v classifier of gender specific words from gender neutral words. Note that it is acceptable to infer *he*:*she* into *brother*:*sister*, but is considered biased to construct a relation between *he*:*she* and *programmer*:*homemaker*. The results show that gender stereotype is prevalent across different word embeddings. Considering the wide range of usage of word embedding in natural language processing, reflections of stereotypes in word embedding not only present but can also amplify the differences in gender rules which can lead to significant issues with machine learning and its applications[7].

2. Motivation

Achieving non-discriminatory machine learning may sound as easy as excluding protected attributes such as age, color, disability, or gender. Contrary to this belief, the study shows that while excluding protected attributes is not enough to remove discrimination in the dataset.

	Training data		Classification result		Classification without sex attribute	
	Male	Female	Male	Female	Male	Female
High income	3256	590	4559	422	4134	567
Low income	7604	4831	6301	4999	6726	4854

Table[10] above shows the number of male and female individuals in high and low income class. In input training data, approximately 30% of male individuals have high income while only about 11% of female individuals have high income. When a naive Bayes classifier learns from this dataset with 24% of

all individuals, the difference in income between sex attribute results in even more distinctions between the male and female individuals with 42% of male individuals with high income and 7.8% of female individuals as shown on classification result column. Even without the sex attribute in the data, the male individual still has a dominant percentage of high income which show favoritism on male individuals by the naive Bayes classifier with 38% of male individuals and 10% of female. This issue is called the *red-lining effect* where the classifier uses attributes that correlate with the discriminatory attribute. In other words, classification with removed direct discrimination attributes would still show discrimination indirectly[10]. Contrary to belief of Kamishima et al. (2011) where the discrimination of income between two sex should be removed[10], Žliobaitė et al. (2011) claims that income gaps between males and females can be justified by considering hours per week (average work hours per week of male works 6 hours longer than those of female individual) and only the part of discrimination which is not explainable by other characteristics should be removed[15]. This paper will discuss so-called conditional discrimination in a later section.

Another issue of measuring discrimination and decision making solely based on statistical parity[2] is that it may cause reverse discrimination. Making employment decisions in favor of minority groups despite the experience or qualification is one of the forms of reverse discrimination and leads to discrimination against members of a majority group. The U.S. Supreme Court endorsed the usage of affirmative action in *Regents of the University of California v. Bakke*(1978) in which a Caucasian medical school applicant challenged a university's decision based on race. The Court held that race could be considered as one of the factors in a college admission policy, but those admission decisions are not allowed to use specific quotas based on race or any other single attributes.

Non-discriminatory learning is a complicated problem which is not only related to categories of classification algorithms, but also highly related to social, economic, technical, and political backgrounds. Fairness in machine learning can be considered as an optimization problem between fairness, accuracy, and data loss[2] in terms of bottleneck[17] or knapsack problem[4]. The goal of bottleneck is to aim the information while preserving information about relevant variables to optimize to find a new representation to maximize the mutual information with X while minimizing the information about Y where we are trying to maximize the individual similarities of the data while minimizing the information with protected attributes such as sex, race, and political background. While a variety of learning algorithms use different similarity measures and discrimination detection, it is important to have an evaluation standard that reflects current society and produces statistically meaningful dataset and output values. This paper will review related and current work in the field of fairness in machine learning and artificial intelligence, and refine evaluation standards for discrimination aware learning algorithms.

3. Related Work

3.1 Direct and Indirect Discrimination

Pedreschi et. al. (2008) define direct and indirect discrimination as direct discrimination consists of rules of procedures that explicitly impose minority or disadvantages group and indirect discrimination as discrimination consists of rules or procedures intentionally or impose the same disproportionate burdens while not explicitly being directly discriminate[1]. Compared to the previous German credit class, which

clearly shows direct discrimination of racial factor, detecting indirect classification is more challenging. Extended from the example, consider the additional classification rule: conf({neighborhood=10451, city =NYC==>bad)=95%. Even though postal code may seem potentially discrimination factors, there are no protected attributes present in this rule stated exclusively without additional data. However, assuming that residents from neighborhood 10451 are dominantly black creates the following association rule: conf({neighborhood =10451, city=NYC==>race=black)=80%. The last can infer and "redline" postal code attribute into potentially distributed attributes. Unlike the direct discrimination in previous examples where they could remove race attributes from the dataset, postal code may contain valuable information such as development level of location(whether the area is city or rural), transportation, and market values if the loan application was for a mortgage. It would be a great loss of information especially for the mortgage and the decision without information would not suffice to make an accurate decision. Fairness aware machine learning is an optimization problem where the algorithm should aim to minimize the data loss while maximizing the similarity without discrimination.

Occasionally even if the data shows significant differences in the statistics on protected attributes(e.g. race) do not have to be removed. There are discrimination in job appilcation when an individual with foreing status has low acceptance employment rate but it is under the greater condition of lack of experience which is a reasonable attribute to consider on job application. Similarly, Žliobaitė et al. (2011) claim that income differences between male and female can be justified since the confidence rule on sex is followed under hours per week where average work hours per week of male is six hours longer than average work hours of female[15]. Contrary to initial definition of indirect discrimination in Pedreschi et. al. (2008) and previous work of fairness in data analysis[1, 2, 3, 9], Žliobaitė et al. suggest that only the part of discrimination which is not explainable by other characteristics should be removed [15] and define those who can be explainable by other characteristics *bad discrimination*. Although average work hours and experience level are reasonable attributes to determine employment decisions, it still remains on social studies to find relations among attributes. For example, it could be inferred that foreign people have disadvantages in employment as the reason why attributes of work experiences (co) relate to their application which the gap will be further divided.

3. 2 Discrimination Measures

Discrimination measure plays a great rule in detection and prevention on both direct and indirect discrimination. Various discrimination measures are required for different types and purposes of machine learning algorithms. Calders et al. (2010) use three different approaches for removing discrimination from a naive Bayes classifier by using a *discrimination score*, also called Calders-Verwer score(CV score)[11], which defined as the difference between the probability of target group and control group. For the data and two classifiers could achieve three *discrimination score* of data, naive Bayes, and naive bayes without sensitive attributes[10]. In perspective to data mining, Pedreschi et al. (2008) introduced the key concept of *elift*(Extended lift) in the notion of association and classification rule. In general, the extended lift range over $[0, \infty)$ without minimum support consideration.

Definition 1. Extended lift. [1]Let $A, B \rightarrow C$ be as association rule such that $conf(B \rightarrow C) > 0$. Extended lift of the rule with respect to B as:

$$\frac{conf(A,B\rightarrow C)}{conf(B\rightarrow C)}$$

Where B is defined as the *context, and* $B \rightarrow C$ is defined as the *base-rule*[1].

The extended lift calculates the impact on confidence due to the additional itemset A in confidence score of the base-rule, $B \rightarrow C$. While the extended lift can be easily calculated based on the given dataset and widely used on discrimination detection, it calculates the confidence score under the assumption that all attributes are independent of each other, which is unlikely to be the case in the real world. Another limitation of solely using the extended lift as a discrimination measure is that it may not be able to distinguish *bad discrimination*[15].

	Medicine		Computer Science	
	Male	Female	Male	Female
Number of applicants	200	800	800	200
Acceptance rate	20%	20%	40%	40%
Accepted(+)	40	160	320	80

Žliobaitė et al. (2011) used the table above to demonstrate the *bad discrimination* which should not be removed. Based on the total number of accepted applicants, this table shows 12% discrimination between male applicants and female applicants. However, when we look closely into this data, it shows that since female applicants favor the medicine program, which is more competitive than computer science, the results show lower total acceptance rates for females than those of male applicants even with the same acceptance rate based on individual programs. The Fall applications of 1973 at the University of California shows a real-life example of the absence of *bad discrimination,* where all 44% of males and 35% of female applicants are admitted with 9% discrimination toward women but the examination in different department shows statistically significant bias in favor of female[16]. Manchuhan and Clifton(2014) used *belift* to discover discrimination. The extended lift can be adapted into the Bayesian network by adapting calculation of the denominator probabilities with Bayesian network[5] and the *belift* is based on the class probabilities from the Bayesian network and the relative Bayesian network based on Bayesian theorem as shown in <Definition 2>. The discrimination algorithm using *belift* returns the Bayesian network, the discriminated instances, and the non-discriminated intances[5].

Definition 2. Belift. [5] Let A be the input variables where A^r denotes redlining attributes of the protected variable. The *beilft* of the rule respect to X^α as:

$$\frac{P(y^+|s^1,X^r,X^\alpha)}{P(y^+|X^\alpha)}$$

Bolukbasi et al. adapt cosine distance between gender direction ($\overline{she} - \overline{he}$ and $\overline{woman} - \overline{man}$) with the subset of 218 gender specific words from w2vNEWS out of 26,377 words in the filtered embedding labeled using crowdsourcing and manual labeling[5]. Zemel et al. (2013) measure discrimination based on the bias with protected attributes in classification as another form of statistical parity with the difference between positively classified individuals in the protected groups and unprotected groups [4].

3. 3 Fairness Algorithms

Non discriminatory decision-making process can be categorized as three general strategies based on how to achieve fairness during the procedures. One involves modifying the labels of input data called *data-massaging strategy* [9, 10]. This strategy *massages* the input training dataset to remove discrimination. Direct rule protection method 1 of Hajian and Domingo-Ferrer (2013) falls under this category where they modify the discriminatory itemset from the database before database data is put into learning algorithms. As shown on line 7-10 in <Algorithm 1>, direct rule protection method 1 relabels and recalculates the confidence score until the confidence score of rule r' is less or equal to $\alpha \cdot conf(B \rightarrow C)$ for direct α-discrimination of each rule r'. This algorithm sorts the database component by ascending impact to minimize the loss of data presented in an initial database and return the new database for training.

Algorithm 1. DIRECT RULE PROTECTION (METHOD 1) [3]

1: **for** each r': $A, B \rightarrow C \in$ {*discrimination rule*} **do**
2: $FR \leftarrow FR - \{r'\}$
3: $DB_c \leftarrow$ All records completely supporting $\neg A$, $B \rightarrow \neg C$
4: Compute $impact(db_c)$
5: **end for**
6: Sort DB_c by ascending impact
7: **while** $conf(r') \geq \alpha \cdot conf(B \rightarrow C)$ **do**
8: Select first record in DB_c
9: Modify discriminatory item set of db_c, from $\neg A$ to A in DB
10: Calculate $conf(r')$
11: **end while**
12: **end for**
13: Output: $DB' = DB$

The second strategy consists of post regularized approaches where leaning algorithms regularize on the resulting decisions [9, 10]. Modifying Naive Bayes reassigns probability until the *discrimination score*(Calders-Verwer score) is greater than 0 by adjusting with the given adjustment rate of 0.01 by modifying the probability distribution.

Algorithm 2. MODIFYING NAIVE BAYES [9]

1: Calculate discrimination score $disc$ in the labels assigned by M to D
2: **while** $disc > 0.0$ **do**
3: $numpos$ is the number of positive labels assigned by M to D
4: **if** $numpos <$ the number of positive labels in D
5: $N(C_+, S_-) = N(C_+, S_-) + 0.01 \times N(C_-, S_+)$
6: $N(C_-, S_-) = N(C_+, S_-) - 0.01 \times N(C_-, S_+)$
7: **else**

8: $N(C_-, S_+) = N(C_-, S_+) + 0.01 \times N(C_+, S_-)$
9: $N(C_+, S_+) = N(C_-, S_+) + 0.01 \times N(C_+, S_-)$
10: **end if**
11: Update M using the modified occurrence counts N for C and S
12: Calculate *disc*
13: **end while**

While <Algorithm 2> can easily achieve zero discrimination on sensitive attribute S in the given class value of S, it is unrealistic to expect the modifying Naive Bayes algorithm to detect a *red-lining effect.* Naive Bayes approaches generally perform well at achieving non-discriminatory machine learning but have difficulty detecting low value of discrimination at test time[9]. The last category of strategy uses designed algorithms to achieve discrimination free classification rather than modifying the dataset or end results of learning algorithms such as "Fairness through Awareness" work[2]. Dwork et al. (2011) obtained intermediate representation using mapping by optimizing the classification decision methods while satisfying Lipschitz condition on individual similarity[2]. Condition L is defined as Lipschitz condition when A and B are similar in L condition under individual similarity metric for task T. Discrimination aware machine learning is not limited to logistic regression and Bayesian method but expanded to various types of machine learning algorithms including top k-measure[20], decision tree[11], and k-NN clasifier[25].

4. Evaluation Standards

Each of machine learning requires different discrimination measures and unbiased algorithms to achieve ideal discrimination-free results based on their purposes and designs. It is important to have standards of evaluation for fairness aware algorithms that can evaluate and compare performance among algorithms. Even though not every learning algorithm can be quantified under every standard(e.g. Accuracy of w2v on gender neutral distance), it is beneficial to have standards for evaluation measure for future references. This paper introduces four standards including accuracy, fairness, consistency, and related attribute set to provide guidelines to evaluate discrimination aware learning algorithms and data mining.

1. ***Accuracy***: Maintaining the accuracy of the model classification prediction and preventing data loss are equally important as achieving fairness.
2. ***Fairness/Discrimination***: Discrimination on protecting attributes including age, colour, disability, gender, national origin, political affirmation, race, religion, sexual orientation, and veteran status are prohibited by laws. The model should be considered the complication of relations among attributes including social, economic, and physical backgrounds.
3. ***Consistency***: Model supposed to preserve the individual similarities before a model's classification prediction. Zemel et al. applied the kNN function on every individual to obtain an accurate representation of each point's relative distances and nearest neighbors[4].
4. ***Related Attribute set***: The new evaluation standards propose to include a related attribute set of a model on evaluation measure. To handle conditional discrimination[15], a model should provide all related decision making attributes not limited to protected attributes and each attribute's factor.

A related attribute set can be advised to review comprehensively with social, economic, and other complicated relations in society.

5. Conclusion and Discussion

This paper explored and compare definitions of discrimination in machine learning and data mining, comparing various machine learning algorithms, and introduced suggestions of evaluation standards to measure non-discriminatory learning algorithms in terms of fairness, accuracy, reflectiveness of attributes, consistency, and data loss. As machine learning influence has become significant in decision making, it is important to measure and prevent discrimination which not only present but also has the potential to amplify the unfairness among minority groups. One cannot express enough of the social and economic studies behind algorithms to provide linkage between attributes and (co)relations which are essential for fairness and accuracy results for future work. It is important to recognize physical, economical, social, and biological backgrounds and reconcile the differences through the true meaning of fairness.

Lastly, it is important to put emphasis on the awareness of gender fluidity and multiethnic placement for the future works in fairness in machine learning. Gender and race play significant roles in individual attributes and identities. PEW Research center analysis of 2014 American Community survey and 1980 and 2000 decennial census(IPUMS) shows the percentage of children younger than 1 who are multiracial or multiethnic, among those living with two parents rate has risen from 1% in 1970 to 15% in 2015. This rate is three times faster than population growth as a whole. As the diversity of the world grows and as the world becomes globalized, learning algorithms should be aware and adapt changes respectively.

References

1. Dino Pedreschi, Salvatore Ruggieri, Franco Turini. Discrimination-Aware Data Mining. Knowledge Discovery and Data Mining, 2008.

2. Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold. Fairness Through Awareness. Innovations in Theoretical Computer Science Conference, 2012.

3. Sara Hajian and Josep Domingo-Ferrer. A Methodology for Direct and Indirect Discrimination Prevention in Data Mining. IEEE Transactions on Knowledge and Data Engineering, 2013.

4. Richard Zemel, Yu Wu, Kevin Swersky, Toniann Pitassi, Cynthia Dwork. Learning Fair Representations. International Conference on Machine Learning, 2013.

5. Koray Mancuhan, Chris Clifton. Combating Discrimination Using Bayesian networks. Artificial Intelligence and Law, 2014.

6. Francesco Bonchi, Sara Hajian, Bud Mishra, Daniele Ramazzotti. Exposing the Probabilistic Causal Structure of Discrimination. ArXiv Preprint, 2015.

7. Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, Adam Kalai. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. Neural Information Processing Systems, 2016.

8. J. Gordon and B. Van Durme. Reporting bias and knowledge Acquisition.

9. Calders, Toon, and Sicco Verwer. “Three Naive Bayes Approaches for Discrimination-Free Classification.” Data Mining and Knowledge Discovery 21, no. 2 (July 27, 2010): 277–292. doi:10.1007/s10618-010-0190-x

10. Kamishima, T. Akaho, S., and Sakuma, J. Fairness aware learning through regularization approach. In *IEEE 11th International Conference on Data Mining,* pp 643-650, 2011

11. Calders T, Kamiran F, Pechenizkiy M (2010) Constructing decision trees under independency constraints. Technical report, TU Eindhoven

12. G.Ausiello,P.Crescenzi,G.Gambosi,V.Kann,A.Marchetti- Spaccamela, and M. Prosati. Complexity and Approximation. Combinatorial Optimization Problems and Their Approxima- bility Properties. Springer, 2003.

13. Hajian, Sara, Josep Domingo-Ferrer, and Oriol Farràs. “Generalization-Based Privacy Preservation and Discrimination Prevention in Data Publishing and Mining.” Data Mining and Knowledge Discovery 28, no. 5-6 (January 25, 2014): 1158-1188. doi:10.1007/s10618-014-0346-1

14. Pedreschi, Dino, Salvatore Ruggieri, and Franco Turini. “Measuring Discrimination in Socially-Sensitive Decision Records.” Proceedings of the 2009 SIAM International Conference on Data Mining. (2009): 581-592. doi:10.1137/1.9781611972795.50

15. Žliobaitė, Indre, Faisal Kamiran, and Toon Calders. “Handling Conditional Discrimination." 2011 IEEE 11th International Conference on Data Mining. (December 11-14, 2011): 992–1001. doi:10.1109/ICDM.2011.72

16. Tishby, N., Pereira, F.C., and Bialek, W. The In- formation Bottleneck method. In The 37th Annual Allerton Conference on Communication, Control, and Computing, 1999.

17. Berendt, Bettina, and Soren Preibusch. “Better Decision Support Through Exploratory Discrimination-Aware Data Mining: Foundations and Empirical Evidence.” Artificial Intelligence and Law 22, no. 2 (January 10, 2014): 175–209. doi:10.1007/s10506-013-9152-0

18. Romei, Andrea, Salvatore Ruggieri, and Franco Turini. “Discovering Gender Discrimination in Project Funding." 2012 IEEE 12th International Conference on Data Mining Workshops. (December 10, 2012): 394–401. doi:10.1109/ICDMW.2012.39

19. Kamishima, Toshihiro, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. “Fairness-Aware Classifier with Prejudice Remover Regularizer.” 2012 European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases. (September 24-28, 2012): 35–50. doi:10.1007/978-3-642-33486-3_3

20. Pedreschi, Dino, Salvatore Ruggieri, and Franco Turini. “A Study of Top-K Measures for Discrimination Discovery.” 2012 Proceedings of the 27th Annual ACM Symposium on Applied Computing. (March 26, 2012) 126–131. doi:10.1145/2245276.2245303

21. Ruggieri, Salvatore. “Data Anonymity Meets Non-Discrimination.” 2013 IEEE 13th International Conference on Data Mining Workshops (December 7-10, 2013): 875–882. doi:10.1109/ICDMW.2013.56

22. Mascetti, Sergio, Annarita Ricci, and Salvatore Ruggieri. “Introduction to Special Issue on Computational Methods for Enforcing Privacy and Fairness in the Knowledge Society.” Artificial Intelligence and Law 22, no. 2 (February 11, 2014): 109–11. doi:10.1007/s10506-014-9153-7

23. DeDeo, Simon. "Wrong Side of the Tracks: Big Data and Protected Categories" (May 27, 2015). arXiv:1412.4643v2

24. El-Arini, Khalid, Ulrich Paquet, Ralf Herbrich, Jurgen Van Gael, and Blaise Agüera y Arcas. “Transparent User Models for Personalization,” 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. (August 8, 2012): 678-686. doi:10.1145/2339530.2339639

25. Luong, Binh Thanh, Salvatore Ruggieri, and Franco Turini. “K-NN as an Implementation of Situation Testing for Discrimination Discovery and Prevention.” 17th ACM SIGKDD

International Conference on Knowledge Discovery and Data Mining. (July 21, 2011): 502–510. doi:10.1145/2020408.2020488